O & kǫzèpmaǵar zoalactanÿ èlèmzǫ

Novák Attila[1,2], Wenszky Nóra[2]

[1]MTA Nyelvtudományi Intézet
1068 Budapest, Benczúr utca 33.
[2] MTA–PPKE Nyelvtechnológiai Kutatócsoport
1083 Budapest, Práter utca 50/a
novak@nytud.hu

Kivonat: Cikkünkben egy olyan magyar számítógépes morfológiát mutatunk be, amelyet kiegészítettünk az ómagyarban és a középmagyarban még létező, de azóta kihalt alaktani szerkezetek leírásával, illetve a szükséges szókinccsel, így alkalmas régi magyar szövegek elemzésére. Az elemzőt két, a Nyelvtudományi Intézetben párhuzamosan futó, ómagyar, illetve középmagyar szövegek feldolgozásával foglalkozó OTKA kutatási projektben használjuk. A morfológia mellett bemutatjuk a szövegek morfoszintaktikai annotálására használt gépi és kézi egyértelműsítő rendszert, valamint az annotált szövegekben való keresést lehetővé tevő korpuszkezelőt.

1 Bevezetés

A Nyelvtudományi Intézet két OTKA projektjének (Magyar generatív történeti szintaxis [OTKA NK78074], valamint Történeti magánéleti korpusz [OTKA 81189]) feladata többek között az ómagyar és a középmagyar időszakból származó szövegeket tartalmazó morfológiailag elemzett, kereshető korpuszok létrehozása. A projektekben a Humor magyar morfológiai elemző [7] olyan kibővített változatát használjuk, amelyet alkalmassá tettünk a nyelvből időközben kihalt alaktani konstrukciókat, toldalékallomorfokat, toldalékmorfémákat, paradigmákat, töveket tartalmazó szavak elemzésére is. Az alábbiakban áttekintjük az elemzőprogram kifejlesztéséhez szükséges lépéseket, a felmerülő problémákat és megoldásukat, valamint a szövegek morfoszintaktikai annotálására használt gépi és kézi egyértelműsítő rendszert és az annotált szövegekben való keresést lehetővé tevő korpuszkezelőt.

2 A szövegek előfeldolgozása

Mindkét szóban forgó projektnek – a középmagyar szövegekkel foglalkozónak kizárólagos – célja, hogy annotált, kereshető korpuszokat hozzon létre. Míg az ómagyar korból főként kódexek maradtak fenn, és a szövegek nagy része fordítás, a középmagyar korpusz elkészítésekor a célkitűzés az élő nyelvhez sokkal közelebb álló források összeválogatása volt. Így ezt a korpuszt perszövegek – közöttük boszor-

kányperek jegyzőkönyvei – és misszilisek, azaz ténylegesen elküldött főúri és jobbágylevelek alkotják. Az utóbbi korpusz esetében az egyes szövegekhez tartozó metaadatok is fontos szerepet játszanak, amelyek lehetővé teszik ezeknek a forrásoknak történeti-szociolingvisztikai szempontú vizsgálatát is.

2.1 Digitalizálás

A korpuszokat alkotó szövegek eredetileg kéziratos formában maradtak fenn, azonban egyik projektnek sem képezte részét kéziratos szövegek feldolgozása: minden esetben nyomtatott szövegkiadásokból dolgoztunk. A szövegek nagy részének az esetében azonban nem állt rendelkezésre digitalizált szövegváltozat. Így az első feladat a szövegek digitalizálása volt, amelyet az esetek többségében OCR alkalmazásával végeztünk el. Különösen az ómagyar időszakból származó szövegek esetében jelentett nehéz feladatot a szokatlan karakterek és mellékjel-kombinációk feldolgozása. Minden egyes szöveghez újra be kellett tanítani az alkalmazott OCR programot, hiszen másmás különleges karakterek szerepeltek bennük. Az automatikusan felismertetett szövegben azonban így is számos hiba maradt, munkatársainknak tehát minden szöveget végig kellett olvasni. Az eredeti, kinyomtatott szöveget és a digitalizált változatot össze kellett hasonlítani és a beviteli hibákat kézzel javítani.

EVANGELIVM SECVNDVM MATTHAEVM
|| [I]
8ra Jeſus criſtus dauid fia abracham fia zuletetenec ꝙnuꝙ (2) Abracham ke · zule iſaakot' Iſaac ke · zule iacobot Jacob ke · zule iudaſt2 ꝙ attafiait (3) Judas ke · zule phaꝛeſt' 2 Zaramot thamaꝛtol' Phares ke · zule Ezromot Ezꝛom ke · zule Aramot (4) Aram ke · zule Aminadabot Aminadab ke · zule Naaſſont

2.2 Normalizálás

A szövegek rendkívül változatos írásképe, az előforduló sokféle dialektus, illetve az átfogott hosszú időszak folyamán bekövetkezett nagymérvű nyelvtörténeti (elsősorban fonológiai) változások miatt az automatikus elemzés egyik feltétele a szövegek írásképi és fonológiai szempontból egységes formára hozása, azaz normalizálása volt. Ez nagyrészt kézzel történt, és a folyamat során a szövegeket tagmondatokra is bontottuk. A projektben nem volt célunk, hogy olyan elemzőt hozzunk létre, amely a korpuszt alkotó eredeti szövegek teljes fonológiai dialektális változatosságát kezeli. Így a normalizálás során az ilyen jellegű különbségeket – például az *ö*-zést – eltüntettük.

höti	után	vallia
hite	után	vallja:
hit	után	vall
N.PxS3	PP	V.S3.Def

szőrö	szalát	sem	fogta	el,
szőre	szálát	sem	fogta	el.
szőr	szál	sem	fog	el
N.PxS3	N.PxS3.Acc	Adv	V.Past.S3.Def	VPfx

Fontos szempont volt azonban az, hogy morfémák a normalizálás folyamán ne tűnjenek el vagy alakuljanak át más morfémákká: például az elbeszélő múltban álló alakokat nem alakítottuk egyszerű múlt időkké stb. A morfémahűség helyes megvalósításához általában alaposan mérlegelnünk kellett az adott korszak ortográfiájának jellegzetességeit. Törekedtünk rá, hogy a korabeli helyesírás bizonytalanságaiból adódó inherens és ténylegesen feloldhatatlan többértelműségeket lehetőleg ne tüntessük el a normalizálás során.

Az egyik jellegzetes többértelműség a korai szövegek magánhangzóhosszúság-jelölésének hiányából, illetve bizonytalanságából és abból a tényből adódott, hogy a határozott tárgyas igeragozás használatának szabályszerűségei az adott időszakban részben különböztek attól, amit a szöveget normalizáló nyelvészek anyanyelvi intuíciója esetleg sugallna. A szövegek egy részében például egyértelműen megfigyelhető, hogy egyenes idézés esetén – ellentétben a mai köznyelvben szokásostól – a *mond* ige határozatlan ragozással is használatos volt.

mondotta	a	Felesegenek
mondta	a	feleségének:
mond	a	feleség
V.Past.S3.Def	Det	N.PxS3.Dat

arra	**mond**	Lovasz	Matyasne
Arra	**mond**	Lovász	Mátyásné:
az	**mond**	Lovász	Mátyásné
N|Pro.Sub	**V.S3**	N	N

Az elbeszélő múltban azonban a *monda* igealak ebben a helyzetben magánhangzóhosszúság-jelölésének bizonytalansága miatt éppoly kevéssé rekonstruálható módon utal az igeragozás határozott vagy határozatlan voltára *(monda ~ mondá)*, mint a *mondtam* alak. A bizonytalanság forrása itt a rag magánhangzója hosszúságának bizonytalanságából fakad, amelyet a normalizált szövegben ilyen esetben a magánhangzó után írt ékezettel jelölünk.

monda	erre	Giöröfy	Janosne,
Monda´	erre	Györöfy	Jánosné:
mond	ez	Györöfy	Jánosné
V.Ipf.S3.Def?	N|Pro.Sub	N	N

én	**mondottam**	néki	
Én	**mondtam**	neki:	
én	**mond**	ő	
N|Pro.S1	**V.Past.S1.Def?**	N|Pro.Dat.S3	

Hasonlóan bizonytalan az igeragozás határozott volta abban az esetben, ha a tárgy birtokos szerkezet, de nincs definit determinánsa. Ebben az esetben a határozott vagy határozatlan igeragozás használata dialektusfüggő. (Az alábbi példákban a *nyavalyáját* determinánsa a szintén dialektusfüggő definitségű *mely*, a többi birtokos tárgy pedig determináns nélküli). A szöveget normalizáló vagy annotáló személy ilyenkor nem vetítheti a saját intuícióját az adott szövegre. Alább az első két példa a szerzők számára agrammatikus, mert a birtokos szerkezet tárgy mellett mindenképp definit igeragozást használnánk. Azonban mivel tudjuk, hogy más dialektusokban ez nem feltétlenül van így, az elbeszélő múltat tartalmazó harmadik szerkezetet inherensen többértelműnek kell tartanunk, nem tudván, hogy melyik dialektusból származik.

mely	**nyavalájáát**	Tormánéra	**gyanétott,**
mely	**nyavalyáját**	Tormánéra	**gyanított,**
a+mely	**nyavalya**	Tormáné	**gyanít**
Det|Pro|Rel	**N.PxS3.Acc**	N.Sub	**V.Past.S3**

hogy	holt	Ember	**koponyáit**	az	Padlason	**tartott**	volna,
hogy	holt	ember	**koponyáit**	a	padláson	**tartott**	volna,
hogy	holt	ember	**koponya**	a	padlás	**tart**	van
C	Adj	N	**N.PxS3.Pl.Acc**	Det	N.Sup	**V.Past.S3**	V.Cond

azulta	**halla**	rosz	**hirét,**	és	nevét,
azolta	**halla´**	rossz	**hírét**	és	nevét,
azolta_az+óta	**hall**	rossz	**hír**	és	név
Adv|Pro	**V.Ipf.S3.Def?**	Adj	**N.PxS3.Acc**	C	N.PxS3.Acc

Hasonló rendszeres többértelműségek jelentkeznek az elöl képzett tövek i-ző birtokos alakjai esetében, ha egyéb rag is van a szó végén (pl. *cselekedetinek*). Ezekben az esetekben még a szövegkörnyezet alapján sem mindig lehet egyértelműen eldönteni, hogy egyes számú vagy többes számú alakról van szó (*cselekedetének* vs. *cselekedeteinek*). Ilyenkor a normalizálás során meghagyjuk az i-ző birtokos alakot, az elemzőt pedig képessé tettük arra, hogy ezeket a szóalakokat úgy is tudja elemezni hogy a számot bizonytalannak jelöli:

csupán	az	Aszony	**cselekedetinek**	tulajdonitotta,
csupán	az	asszony	**cselekedetinek**	tulajdonította.
csupán	az	asszony	**cselekedet**	tulajdonít
Adv	Det	N	**N.PxS3.Pl?=i.Dat**	V.Past.S3.Def

Egyes szövegek korábbi normalizálása nem az általunk lefektetett elvek szerint történt, ilyen volt pl. a Székelyudvarhelyi kódex. Ennek szövege a mai magyar helyesírásnak megfelelő hangjelölést alkalmaz, azonban a szöveg fonológiai-dialektális sajátosságait nem közelítették a mai magyarhoz, ezért további kézi adaptációra volt szükség.

2.3 A -bAn/bA probléma

A normalizálás és a különösen a morfémahűség megítélése szempontjából speciális problémát jelentett a -bAn, illetve -bA ragos szóalakok kezelése. A két korpusz szövegeinek vizsgálata egyértelműen azt jelzi, hogy a két ragnak a beszélt nyelvben jelenleg sem éles szétválása sok száz éve stabilan fennálló állapot [6] (nevezetesen, hogy a -bA változat szóban minden további nélkül használható a -bAn funkciójában is, miközben az utóbbi változat is létezik és használatos), amely a leírt szövegekben általában meglehetősen zavaros képhez vezetett. A korpusz szövegei egyértelműen jelentősen különböznek abból a szempontból, hogy a feltételezhetően inesszívusz, illetve illatívusz funkciójú elemek jelölésére mennyire következetesen melyik ragalakot írták le. A -bAn/-bA elemeket tartalmazó szóalakok ortográfiája szempontjából merőben különböző megoldásokat találunk a korpuszban, még két egymással apa–fia relációban álló személy (Nádasdy Tamás és Nádasdy Ferenc) esetében is (az előbbi szinte kizárólag a -bA alakot, az utóbbi szinte kizárólag a –bAn-t használja minden funkcióban).

Azért, hogy biztosan ne essünk se abba a hibába, hogy egy merőben ortográfiai ügyet grammatikainak hiszünk, és így hibás elemzések tömkelegét állítjuk elő, se abba, hogy visszakövethetetlen módon mindent átírunk a saját kompetenciánknak

megfelelő alakra, azt a megoldást választottuk, hogy a -bAn/-bA elemeket tartalmazó szóalakok normalizálása során explicite jelöltük azokat az eseteket, ahol mindent a lehető leggondosabban mérlegelve úgy ítéltük, hogy a leírt alak nem felel meg a szándékolt grammatikai funkciónak, illetve az általunk használt ortográfiai normának, így a normalizált alak és az elemzés alapján visszakereshetők és kvantifikálhatók az egyes szövegek a -bAn/-bA-jellemzői.

az	Macska	az	Tehent	szopja	az	olba
A	macska	a	tehént	szopja	az	ólba'."
a	macska	a	tehén	szop	az	ól
Det	N	Det	N.Acc	V.S3.Def	Det	N.Ine

azomba	Bekene	aszt	mondotta
Azonba'	Bekéné	azt	mondta:
azonban	Bekéné	az	mond
C	N	N|Pro.Acc	V.Past.S3.Def

az	asztal	melől,	a	tüz	eleiben	ment,
Az	asztal	mellől	a	tűz	eleibe'n	ment,
az	asztal	mellől	a	tűz	eleibe_elébe	megy
Det	N	PP	Det	N	PP	V.Past.S3

azomban	midőn	bé ment	az	Orvos	Házaban
azonban	midőn	bement	az	orvos	házába'n,
azonban	a+midőn	be|+megy	az	orvos	ház
C	Adv|Pro|Rel	V.Past.S3	Det	N	N.PxS3.III

2.3 Jakab-féle adattárak

Az ómagyar kódexek egy része (a Jókai- [2], a Guary- [3], az Apor- [4] és a Festetics-kódex [5]) szótárszerű formában számítógépes nyelvtörténeti adattárként Jakab László debreceni kollektívája által feldolgozva volt elérhető. Ezekből az 1978 és 2002 között készült kiadásokból igen komoly erőfeszítést igényelt a szövegek visszaállítása. Bár ezek kézzel készült elemzést tartalmaztak, az nehezen olvasható numerikus kódok formájában szerepelt. Az olvashatatlan reprezentációból következő módon gyakran hibás, hiányos, ezen kívül – elsősorban a zárt szóosztályok elemei esetében – az általunk használt elemzésekkel inkompatibilis volt. Ennek ellenére sikerült a szövegeket a szótárakból visszaállítani, az elemzéseket konvertálni és kiegészíteni, ezek alapján automatikusan normalizált változatot generálni, és azt újraelemezni.

A Jakab-féle szótárszerű kiadásokban a szavak az eredeti kódexbeli előfordulásuk helyét (locusát) az oldal/kolumna és az azon belüli sorszám szintjén adták meg. Az alábbi részlet a Jókai-kódex szótárkiadásából származik.

080/08	**ablak**	ablakba	0002	000000	02	11	000	00	05	01
180/15	ablak	ablakbalol	0002	000000	02	11	000	00	09	01
109/12	ablak	ablakokba	0002	000000	02	11	000	01	05	01
159/03	ablak	ablakarol	0000	000000	02	11	000	13	17	01
126/08	**ábráz**	abraz	0000	000000	02	41	000	00	00	01
125/26	ábráz	abrazban	0000	000000	02	41	000	00	08	01
130/22	**abrosz**	Abroz	0000	000000	02	11	000	00	00	01
083/20	abrosz	abrozokott	0003	200000	02	11	000	01	01	01
034/24	**ad**	ad	0000	000000	01	11	000	00	06	01
062/15	ad	ad	0000	000000	01	11	000	00	06	01
082/19	ad	ad	0000	000000	01	11	000	00	06	01

A gyakori szavaknak nem minden előfordulása szerepel ténylegesen a szótári részben. Egy külön függelékben elemzés nélkül fel vannak sorolva az egyéb előfordulá-

sok és írásváltozatok, amelyek közül szerencsés esetben az egyiknél az elemzés is megtalálható. A függelék formája következményeként egyetlen hiba szóelőfordulások tucatjainak rossz elemzését eredményezhette, és eredményezte is.

UTÁN ~ UTÁNA

8/6, 38/8, 63/3, 101/13, 105/14, 106/1, 107/1, 122/7, 132/20, 143/27, 156/7, *vtan* 14/22, 24/25, 62/8, 99/16, 109/26, 120/1, 122/14, 160/26, *vttan* 143/8 *(20 adat)*
18/22, 22/24, 76/17, 90/2, 98/6, 101/8, 106/24, 130/7, 148/10, 160/26, *uttanna* 39/13, 79/14, 132/14, *uta[n]na* 38/22, 101/14, *vtanna* 7/25, 15/17, 25/23, 24, 51/17, 78/10, 138/14, 144/26, 150/16, *vttanna* 57/23 *(25 adat)*

(Összesen: 45 adat)

Az egyes sorok szavainak sorrendjét kézzel kellett a nyomtatott kiadás segítségével helyreállítani. A munkát némileg nehezítette, hogy ugyanabban a sorban néha többször szerepelt ugyanaz a szó – esetleg különböző elemzéssel, de ezekben az esetekben a szótárban általában csak egy előfordulás volt megadva.

003/15	mond	**Monda**	0	1	**11**	1	**13**	**0**	1	0
003/15	ön	ewn	0	6	11	200	0	4	1	0
003/16	jonh	yonhanban	0	2	21	0	13	8	1	3
005/17	s	s	0	10	11	0	0	0	0	0
005/17	mond	**monda**	0	1	**11**	1	**10**	**6**	1	0
005/17	atyjafia	**Attyamfÿa**	100	2	**12**	0	**13**	**0**	3	9
005/18	Ferenc	ferenc	0	3	11	0	0	0	1	0
006/10	de1	De	0	10	11	0	0	0	0	0
006/10	úr	vr	0	2	11	0	0	0	2	0
006/10	Bernald	bernald	0	3	21	0	0	0	1	0
006/10	mond	**monda**	0	1	**11**	1	**12**	**20**	1	3

A visszaállított szövegek számkódos morfológiai elemzéseit programmal konvertáltuk olvasható – és amennyire lehetséges volt – az időközben elkészült morfológiai elemző címkéivel kompatibilis elemzésekké. Ezekre az elemzésekre a morfológiát generátorként alkalmazva megkaptuk a szavak normalizált alakját is.

Ezeket az eredeti szóalakokkal összevetve alább világosan látszanak azok az esetek, ahol a szótárkiadásban hibás elemzés szerepelt, vagy esetleg a feldolgozás során került valamilyen hibás adat az anyagba. Alább az 5/17 *atyámfia* helyett az *atyjafia*, illetve a 6/10 *mondá* vagy *monda* (ez éppen a korábban említett kérdéses definitségű szóalak) helyett a *mondám* szóalak elemzése – ez a hiba a szóalak gyakorisága folytán a szótár függelékében megadott hivatkozás hibás feloldása miatt 106 szóalakot érintett a Jókai-kódexben. Szerencsére ez a hiba könnyen javítható volt.

003 15	Monda	**mondá**	**mond[V.Ipf.S3.Def]**
003 15	ewn	ön	ön[N|Pro.Nom_gen]
003 16	yonhanban	jonhában	jonh[N.PxS3.Ine]
005 17	s	s	s[C]
005 17	monda	**monda**	**mond[V.Ipf.S3]**
005 17	Attyamfÿa	**atyjafia**	**atyjafia[N.PxS3]**
005 18	ferenc	Ferenc.	Ferenc[N]
006 10	De	de	de[C]

006 10	vr	úr	úr[N]
006 10	bernald	Bernald	Bernald[N]
006 10	monda	**mondám**	**mond[V.Ipf.S1.Def]**

A kigenerált szóalakokat eztán újraelemeztük, mert az adattárban megadott elemzések egy része hiányos, illetve az elemző által visszaadott elemzésekkel inkompatibilis volt (elsősorban a névmások és az igenevek esetében). A kapott elemzések közül az adattárban megadotthoz leghasonlóbbat választottuk. Az alkalmazott hasonlósági mérték a trigramhasonlóság volt, amelyet meghatározott heurisztikus konverziók után alkalmaztunk.

A Jakab-féle kódrendszer legsúlyosabb hiányossága az volt, hogy az igenevek fajtáit és ragozott alakjait az általuk használt kódrendszer nem különböztette meg. Ezért ezeket a szavakat és a valódi elemzésüket a program az eredeti ómagyar írásmódú szóalakot is figyelembe véve különböző heurisztikákra alapozva próbálta rekonstruálni. Az alábbi tagmondatban például három szóalak (*p[ro]phetalo, vilagossolot, lattuan*) is igenévként szerepel (14-es kód), de semmi egyéb információ nem derül ki a kódokból sem az igenév fajtájára, sem az esetleges további ragokra vonatkozólag.

005/02	de	De	0	10	11	0	0	0	0	0
005/02	prófétál	**p[ro]phetalo**	0	**14**	11	120	**0**	**0**	10	100
005/02	lélek	lelekuel	4000	2	11	2	0	19	4	0
005/03	világosul	**vilagossolot**	100302	**14**	21	522	**0**	**0**	1	1
005/03	eleve	eleue	0	7	11	0	0	29	0	5
005/03	lát	**lattuan**	0	**14**	11	20	**0**	**0**	0	5
005/03	nagy	nagÿ	0	7	31	0	0	0	0	0
005/03	gond	gondokot	200000	2	11	0	1	1	1	0

A szövegen a fent leírt transzformációkat alkalmazva az alábbit kaptuk:

005	02	De	de	de[C]
005	02	p{ro}phetalo	**prófétáló**	**prófétál[V.PartPrs]**
005	02	lelekuel	lélekkel	lélek[N.Ins]
005	03	vilagossolot	**világosult**	**világosul[V.PartPrf]**
005	03	eleue==	eleve	eleve[Adv]
005	03	lattuan	**látván**	**lát[V.PartAdv=vÁn]**
005	03	nagÿ	nagy	nagy[Adv]
005	03	gondokot	gondokat	gond[N.Pl.Acc]

Az így automatikusan generált szöveget ezután még kézzel ellenőrizni kellett.

3 A morfológiai elemző

A digitalizált és normalizált szövegek elemzésére a Humor magyar morfológiai elemző [7] egy erre a célra kibővített változatát alkalmaztuk. Ehhez ki kellett bővíteni a program tőtárát és toldaléktárát az időközben kihalt paradigmákkal, szótövekkel és toldalékokkal, illetve toldalékallomorfokkal. Az alábbiakban az utóbbiakra láthatunk példákat (félkövérrel kiemelve).

sok	ỹdeyglen	ezkeppen	Asisia	altal	menuen
sok	**ideiglen**	ekképpen	Assisia	által	menvén
sok	**idő**	ekképpen	Assisia	által	megy
Q	**N.Ter=iglAn**	Adv|Pro	N	PP	V.PartAdv=vÁn

Es	yme	efelkoron	kezde	az	hews	asoytanỹa
és	íme	**éjfélkoron**	kezde	az	hős	ásítania,
és	íme	**éjfél**	kezd	az	hős	ásít
C	Int/Utt	**N.Tem=koron**	V.Ipf.S3	Det	N	V.Inf.S3

hoǵ ha	ki	ne{m}	monthatnaia	az	oldalann	is	ki	facadna
hogyha	ki	nem	**mondhatnája**	az	oldalán	is	ki	fakadna.
hogy+ha	ki	nem	**mond**	az	oldal	is	ki	fakad
C	VPfx	Adv	**V.Mod.Cond.S3=nÁjA.Def**	Det	N.PxS3.Sup	C	VPfx	V.Cond.S3

Es	zent	ferencznek	alazatost	engeduen
és	szent	Ferencnek	**alázatost**	engedvén
és	szent	Ferenc	**alázatos**	enged
C	Adj	N.Dat_gen	**Adj.Essmod=t**	V.PartAdv=vÁn

mert	ha	mindọn	iozagoncat	oztogatnoc	istenert	zegeńọknec
mert	ha	minden	jószágunkat	**osztogatnók**	istenért	szegényeknek
mert	ha	minden	jószág	**osztogat**	isten	szegény
C	C	N|Pro	N.PxP1.Acc	**V.Cond.P1=nÓk.Def**	N.Cau	N.Pl.Dat

A elemző toldaléktárába 50 új toldalékot vettünk fel (ezek alakváltozatait, allomorfjait nem számolva). Az alábbiakban olyan toldalékmorfémákra láthatunk példákat az igei morfológia köréből (félkövérrel kiemelve), amelyek a mai magyarban már nem használatosak.

Es	marada	a	nẽbẽṙi	ọ	kèt	fiaytol	&	ferietọl	meǵvalt·
és	**marada**	a	némber	ő	két	fiaitól	és	férjétől	megvált.
és	**marad**	a	némber	ő	két	fiú	és	férj	meg|+válik
C	**V.Ipf.S3**	Det	N	N|Pro.S3	Q	N.PxS3.Pl.Abl	C	N.PxS3.Abl	V.PartPrf

Azért	èlmènè	ember	ludanac	Bètlèhèmé~~bèn~~bọl
azért	**elmene**	ember	Júdeának	Betleheméből,
azért	**el|+megy**	ember	Júdea	betlehem
Adv|Pro	**V.Ipf.S3**	N	N.Dat	N.PxS3.Ela

lọn~~[èf]~~	èhſéǵ	a	fọldọ
lőn	éhség	a	földön.
lesz	éhség	a	föld
V.Ipf.S3	N	Det	N.Sup

bátyádat	oda	affelé	bocsátók	két ezer	emberrel.
Bátyádat	oda,	afelé	**bocsátók**	kétezer	emberrel,
báty	oda	a|+felé+	**bocsát**	két+ezer	ember
N.PxS2.Acc	Adv|Pro	Adv|Pro	**V.Ipf.P1.Def**	Q	N.Ins

Es	valobizoń	azt	nem	foghatoc	meg
és	valóbizony	azt	nem	**foghatók**	meg,
és	való+bizony	az	nem	**fog**	meg
C	Adv	N|Pro.Acc	Adv	**V.Mod.Ipf.P1.Def**	C

A me'l	fọld	tègedèt	m̃ghalauad	fogadād
amely	föld	tégedet	**meghalvád**	**fogadand,**
a+mely	föld	te	**meg|+hal**	**fogad**
N|Pro|Rel	N	N|Pro.S2.Acc	**V.PartAdv.S2**	**V.Fut.S3**

Bizonyos toldalékok produktivitása jelentősen csökkent az ómagyar korhoz képest. Bár ezekkel alkotott szavak a mai magyarban is léteznek, de csak lexikalizálódott egységekként, gyakran az eredetitől különböző szemantikával. Ilyen például az *-At*

képző, ami eredetileg a nomen actionis képző szerepét töltötte be, és teljesen produktív volt. Ennek szerepét vette át később az *-Ás* képző. Jelenleg a cselekvés tárgyi eredményét jelöli (nomen facti, pl. *épület, falazat*) – már ha a szó egyáltalán létezik.

Arra vonatkozólag, hogy az egyes toldalékoknak mely alakváltozatai a töveknek mely alakváltozataihoz kapcsolódtak, tehát hogyan alakultak a paradigmák, nemigen találtunk jól használható leírást. Az adatokat sokszor magukból a forrásokból kellett kideríteni. Bizonyos, időközben kihalt alaktani konstrukciókra viszonylag kevés adat van (pl. az alábbi egyeztetett határozói igenevekre), ráadásul a paradigmák számos elemére sokszor van egyéb lehetséges elemzés is. Ezek formális leírása ezért néha komoly kihívást jelentett.

èl	kẹd	mv́	vrõc	iſtènọ̃c	mèt	ọ	Angala	mègọrizọt	ènget	innèt
él	kedig	mi	Urunk	Istenünk,	mert	ő	angyala	megőrzött	engem	innét
él	kedig	mi	Ur	isten	mert	ő	angyal	meg|+őriz	én	innét
V.S3	C	N|Pro.P1	N.PxP1	N.PxP1	C	N.Nom_gen	N.PxS3	V.Past.S3	N|Pro.S1.Acc	Adv|Pro

èlmènèttèm	es ·	ot	**lakattam**	es ·	&	onnat	idè	fordolattam
elmenettem	is,	ott	**lakattam**	is,	és	onnan	ide	fordulattam.
el|+megy	is	ott	**lakik**	is	és	onnan	ide	fordul
VPfx.V.PartAdv=AttA.S1	Adv	Adv|Pro	**V.PartAdv=AttA.S1**	Adv	C	Adv|Pro	Adv|Pro	V.PartAdv=AttA.S1

lők	lől	lőn
lesz[V.Ipf.S1]	lesz[V.Ipf.S2]	lesz[V.Ipf.S3]
lőnk	lőtök	lőnek
lesz[V.Ipf.P1]	lesz[V.Ipf.P2]	lesz[V.Ipf.P3]

fekvém	fekvéd	fekvén
fekszik[V.PartAdv.S1]	fekszik[V.PartAdv.S2]	fekszik[V.PartAdv=vÁn]
fekvénk	fekvétek	fekvéjük
fekszik[V.PartAdv.P1]	fekszik[V.PartAdv.P2]	fekszik[V.PartAdv.P3]

A toldalékok és paradigmák leírásánál nagyságrendileg több munkát jelentett azoknak a töveknek a felvétele, amelyek a mai magyar elemző lexikonából hiányoztak. Sok esetben a tő ugyan megvolt, de a régi szövegekben más szófajú (is) volt, mint ma, illetve bizonyos konstrukciókban másképp kell elemezni őket, mint a mai megfelelőjüket. Ilyen például a régi névutós szerkezetek egy része, amelyben a névutó a -nAk-os birtokos szerkezethez hasonló formában egyeztetve van az NP fejével, ebben a ragos névutó elemzése más, mint az azonos alakú, ma is létező inkorporált névmást tartalmazó alaké. Kiemelkedően sok munkát jelentett a névmási elemet tartalmazó egységek paradigmáinak szabályszerű leírása.

az	**lövésnek**	**miátta**	oly	nehezen	nem	volna,	gyermeke
a	**lövésnek**	**miatta**	oly	nehezen	nem	volna	gyermeke,
a	**lövés**	**miatt**	oly	nehéz	nem	van	gyermek
Det	**N.Dat**	**PP.PxS3**	Adj|Pro	Adj.Essmod	Adv	V.Cond.S3	N.PxS3

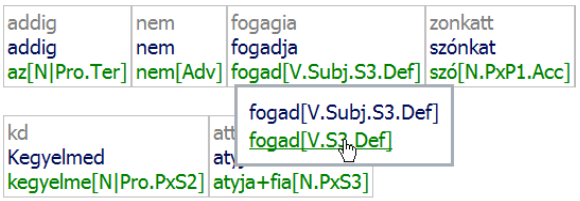
hogy	majd	megholt	**miatta.**
hogy	majd	meghalt	**miatta.**
hogy	majd	meg|+hal	**|+miatt**
C	Adv	VPfx.V.Past.S3	**PP.S3**

4 Egyértelműsítés

A néhány eleve elemzett formában meglévő szövegtől eltekintve a szövegek elemzését egyértelműsíteni is kellett. A lazább, megengedőbb elemző és a kibővített igei paradigmákban szereplő sok egybeesés, valamint a feljebb leírt eldönthetetlen többértelműségek ilyenként való címkézése miatt a történeti szövegekben a többértelműség aránya magasabb, mint a mai szövegek standard Humor elemzővel való elemzése esetében.

A morfoszintaktikai annotáció egyértelműsítésében a munka oroszlánrészét géppel végeztük. Az ó- és középmagyar elemző elemzéseit felhasználva eleinte a HMM-alapú HunPos taggert [1], később a PurePos taggert [8] inkrementális módon egyre több egyértelműsített és ellenőrzött szöveggel betanítva. Mivel a HunPos tövet nem ad vissza, csak címkét, a Humor elemzései közül a HunPos által választotthoz leghasonlóbb címkét tartalmazó elemzést választottuk. A PurePos esetében egyszerűbb a helyzet, mert ezt a feladatot saját hatáskörben elvégzi.

Az így egyértelműsített szövegek kézi ellenőrzéséhez (illetve az első szövegek még teljesen manuális egyértelműsítéséhez) olyan webes felületet hoztunk létre, amelyen a téves egyértelműsítések, illetve normalizálási hibák nagyon hatékonyan javíthatók. Az automatikusan választott elemzés helyett másikat az egérmutatót a szó fölé húzva automatikusan megjelenő listából választva lehet megadni. Kézzel is javítható akár az eredeti, akár a normalizált szóalak, akár az elemzés. A javítás után a szó azonnal újraelemeztethető, és új elemzés választható.



Az elemzőrendszert úgy alakítottuk ki, hogy alkalmas legyen arra, hogy a projekt során az alkalmazott annotáció egyes részleteit meg lehessen változtatni úgy, hogy ugyanakkor ne kelljen kidobni a korábban elvégzett egyértelműsítési munkát, hanem a korábban egyértelműsített szövegekbe is viszonylag egyszerűen átkerüljenek a módosított annotációk. Ennek alapjául az szolgál, hogy a szövegek újraelemzésekor a rendszer automatikusan a korábban megadott elemzéshez leghasonlóbb elemzést választja (az elemzésekből betűhármas-statisztikát készítve, és ezeket összehasonlítva). Bizonyos, az elemzőn végzett változtatások esetében (pl. amikor úgy döntöttünk, hogy a képzett igei alakoknak a korábbiaknál részletesebb elemzését használjuk) ennél kifinomultabb mechanizmusra volt szükség: a már meglevő egyértelműsített elemzéseket géppel generált reguláris kifejezésekkel konvertáltuk.

5 Keresés a korpuszban

A szövegekben való keresést támogató korpuszkezelő nemcsak azt teszi lehetővé, hogy különböző grammatikai szerkezetekre keressünk a szövegekben példákat, hanem azt is, hogy a kereső találataiban is azonnal kijavíthassuk az esetlegesen még az annotációban vagy a szövegben maradt hibákat, amely javítások ilyenkor az adatbázisba azonnal visszakerülnek. (A kereső utóbbi változata csak a megfelelő szakértelemmel és jogosultságokkal rendelkező felhasználók számára elérhető.) A hibakeresés és –javítás egyik hatékony módja, amikor a korpuszban kifejezetten olyan szerkezeteket keresünk, amelyek valószínűleg hibásak, és a valóban hibás találatokat azonnal javítjuk. A javított korpuszt ezután exportálni lehet, és a taggert a javított korpusszal újratanítani.

A keresőrendszer által használt korpuszadatbázis az Emdros korpuszkezelőn alapul [9]. A középmagyar korpusz lekérdezésére használható keresőben az Emdros eredeti lekérdezőszintaxisának (MQL) megfelelően megfogalmazott kérdések mellett egy az MQL-nél jóval tömörebb lekérdezőnyelv is használható. Az utóbbi formában megfogalmazott keresőkérdéseket a rendszer automatikusan MQL-re fordítja.

A kereső lehetővé teszi, hogy mondaton, tagmondaton, vagy adott metaadatokkal megjelölt tulajdonságú szövegen belül keressünk, illetve akár többmondatos egységek is lekérdezhetők. A kereső által megjelenített találati egység a mondat. A tagmondatok lehetnek nem folytonosak (ez az alárendelő szerkezetek esetén gyakran előfordul, de olykor a főmondat vagy egy mellérendelő szerkezet valamelyik eleme ékelődik be). Az alábbi példa olyan találati mondatot mutat be, amelyben több megszakított tagmondat is szerepel.

6 Összefoglalás

Cikkünkben egy ó- és középmagyar szövegek elemzésére is használható számítógépes morfológia kifejlesztésének legfontosabb lépéseit és az eközben felmerülő problémákat és megoldásukat mutattuk be. Emellett bemutattuk azt a keresőrendszert is, amely

lehetővé teszi az annotált szövegekben való keresés mellett azt is, hogy a keresés során kiderülő hibákat az erre jogosult felhasználók azonnal javítsák.

Amellett, hogy sikerült egy megbízhatóan működő, könnyen javítható elemzőprogramot és ennek felhasználásával morfológiailag elemzett történeti korpuszokat létrehozni, a projekt más tanulságokkal is bírt. A bA~bAn végződések speciális kódolása lehetővé tette, hogy a rag ingadozó helyesírásának változásáról számot adjunk. A történeti távlatokban létező szintaktikai többértelműségek néhány körét sikerült jól meghatározni és ezek kódolására, s ezáltal detektálására is sikerült módszert találnunk.

Az elkészült elemzővel a folyamatosan bővített ómagyar és középmagyar korpuszt elemezzük. Az elemzett adatbázisok kereshető formában részben már elérhetők. Az ómagyar korpusz itt: http://rmk.nytud.hu, a középmagyar korpusz feldolgozott része pedig ezen a címen: http://clara.nytud.hu/tmk.

Hivatkozások

1. Halácsy, P., Kornai, A, Oravecz, Cs.: HunPos: an open source trigram tagger. In: Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions (2007) 209–212
2. Jakab L.. A Jókai-kódex mint nyelvi emlék szótárszerű feldolgozásban (Számítógépes nyelvtörténeti adattár 10.). Debreceni Egyetem, Debrecen (2002)
3. Jakab L., Kiss A.: A Guary-kódex ábécérendes adattára (Számítógépes nyelvtörténeti adattár 6.). Debreceni Egyetem, Debrecen (1994)
4. Jakab L., Kiss A.: Az Apor-kódex ábécérendes adattára (Számítógépes nyelvtörténeti adattár 7.). Debreceni Egyetem, Debrecen (1997)
5. Jakab L., Kiss A.: A Festetics-kódex ábécérendes adattára (Számítógépes nyelvtörténeti adattár 9.), Debreceni Egyetem, Debrecen (2001)
6. Németh M.: Nyelvi változás és váltakozás a műveltségi tényezők tükrében. Nyelvi változók a XVIII. században. Szegedi Tudományegyetem, Szeged (2008)
7. Novák A.: Milyen a jó humor? In: Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2003). Szegedi Tudományegyetem, Szeged (2003) 138–145
8. Orosz, Gy., Novák, A.: PurePos – an open source morphological disambiguator. In: Proceedings of the 9th International Workshop on Natural Language Processing and Cognitive Science. Wrocław, Poland (2012)
9. Petersen, U.: Emdros – A Text Database Engine for Analyzed or Annotated Text. In: Proceedings of the 20th International Conference on Computational Linguistics, Volume II. Geneva (2004) 1190–1193